A Loan Risk Prediction Model Using Convolutional Neural Networks and Stacking Fusion

Report submitted to the SASTRA Deemed to be University as the requirement for the course

INT424: ALGORITHMIC TRADING

Submitted by

NANDHINI DEVI S

(Reg.no 125150033)

May 2024

SCHOOL OF ARTS SCIENCE AND HUMANITIES

THANJAVUR, TAMIL NADU, INDIA – 613 401

SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act,1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION
THANJAVUR – 613 401

Bonafide Certificate

This is to certify that the report titled “ **A Loan Risk Prediction Model Using Convolutional Neural Networks and Stacking Fusion**” submitted as a requirement for the course, **INT424: ALGORITHMIC TRADING** for M.Sc. Data Science programme, is a bona fide record of the work done by **Ms.NANDHINI DEVI S(Reg. No.125150033**) during the academic year 2023-24, in the School of Arts, Sciences, Humanities and Education, under my supervision.

Signature of Project Supervisor :

Name with Affiliation : **Ashok Palaniappan**

Date : **01/05/2024**

Project *Viva voc*e held on _01-May-2024

Examiner 1 **Examiner 2**

SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act,1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION

THANJAVUR – 613 401

Declaration

I declare that the report titled "**A Loan Risk Prediction Model Using Convolutional Neural Networks and Stacking Fusion**" submitted by me/us is an original work done by me/us under the guidance of **Dr Ashok Palaniappan, Associate Professor, School of Chemical and Biotechnology, SASTRA Deemed to be University** during the second semester of the academic year 2023-24, in the **School of Arts Science, Humanities And Education**. The work is original and wherever I have used materials from other sources, I have given due credit and cited them in the text of the report. This report has not formed the basis for the award of any degree, diploma, associate-ship, fellowship or other similar title to any candidate of any University.

Signature of the candidate(s) : S. Nandhinidevi

Name of the candidate(s) : **NANDHINI DEVI S**

Date : **01.05.2024**

Acknowledgements

I would like to express my sincere gratitude to Dr Ashok Palaniappan, Ph.D (Illinois), whose invaluable insights and expertise have been instrumental in shaping the direction of this study. The mentorship and encouragement have been a constant source of inspiration throughout the research process. The unwavering support and encouragement have been crucial in navigating the complexities of this field. Furthermore, I would like to thank Sastra Deemed To be University for providing access to research resources and facilities, which have been indispensable in conducting empirical analyses and data-driven experiments. Last but not least, I am deeply grateful to my family and friends for their unwavering support and encouragement throughout this journey. Their patience, understanding, and encouragement have been a source of strength during challenging times.

This research paper is a testament to the collaborative efforts of all those who have contributed to its realization. Thank you for your invaluable support and encouragement.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Several machine learning and deep learning models are implemented and evaluated for a loan risk prediction task using the provided code. The dataset that was employed includes data on applicant and co-applicant income, property area, education, self-employment status, gender, marriage status, dependents, and credit history. Predicting whether or not a loan will be accepted is the aim.Convolutional neural networks (CNNs), artificial neural networks (ANNs), logistic regression as the final estimator in a stacking classifier (Stacking+CNNs), support vector machines (SVMs), K-Nearest Neighbors (KNNs), logistic regression, AdaBoost, and naive bayes are some of the models that were used. Metrics including precision, recall, ROC AUC, F1 score, accuracy, and precision are used to train and assess these models.

The CNN and ANN models are appropriate for examining the features of the dataset because they are specifically developed for handling sequential data. To increase prediction accuracy, the CNN and other models' predictions are combined by the Stacking+CNN model. Other models for comparison include the SVM, KNN, Logistic Regression, AdaBoost, and Naive Bayes models. The efficacy of each model in forecasting loan acceptance is thoroughly examined through the application of the aforementioned indicators to assess its performance. The best model for the loan risk prediction task is determined by comparing the models using visualizations like bar and line graphs, which show performance measures.

CHAPTER 1

INTRODUCTION

1.1 Algorithmic Trading

Using computer algorithms to carry out trading methods is known as algorithmic trading, or "algo" trading. In order to conduct trades, these algorithms adhere to a predetermined set of guidelines. Their goals are to maximize market opportunities, minimize transaction costs, and achieve optimal execution pricing.

The capacity of algorithmic trading to execute trades at frequencies and speeds faster than those of a human being is one of its main advantages. This makes it possible to quickly implement trading strategies in reaction to changes in the market, breaking news, or other variables.Different types of algorithmic trading strategies exist, such as market-making strategies that continuously quote bid and ask prices to provide liquidity to the market, trend-following strategies that seek to profit from market trends, and arbitrage strategies that take advantage of price differences between markets.

Algorithmic trading raises questions about market manipulation, system breakdowns, and the possible impact of automated trading on market stability, even if it can also have benefits like increased efficiency and lower trading costs. Regulatory agencies thus keep a careful eye on algorithmic trading operations to guarantee honest and well-functioning markets.

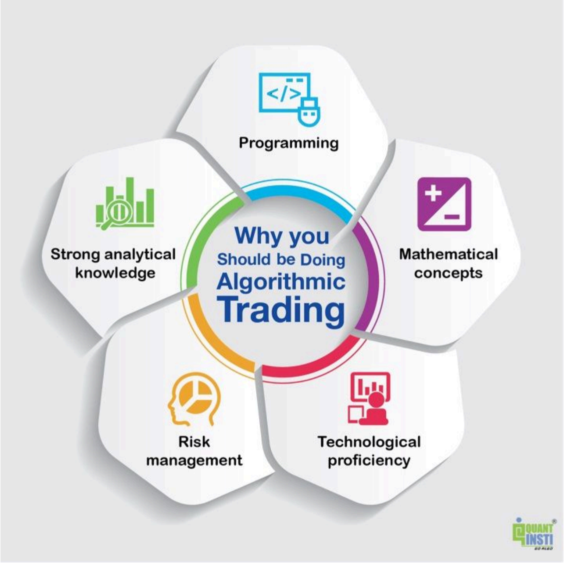

Programming
Strong analytical knowledge
Why you Should be Doing Algorithmic Trading
Mathematical concepts
Risk management
Technological proficiency
QUANT INSTI

Fig 1.1: Algorithm Trading

1.2. About The Loan Risk

The practice of estimating the probability that a borrower would default on a loan by utilizing data and statistical algorithms is known as loan risk prediction. For financial institutions, this is an essential responsibility since it helps them control their risk exposure and make well-informed lending decisions.

Many characteristics or factors pertaining to the borrower and the loan are usually taken into account in order to anticipate loan risk. These could consist of the borrower's income, employment position, credit score, loan amount, loan period, and other pertinent details. Machine learning models that estimate the likelihood of default for new loan applications are trained on historical data on loan performance.

For predicting loan risk, machine learning techniques like logistic regression, random forests, gradient boosting machines, and neural networks are frequently employed. Large data sets can be analyzed by these algorithms to find links and trends that can be used to forecast loan outcomes.

Enhancing the accuracy of loan approval decisions, lowering the default risk, and eventually boosting the profitability and long-term viability of lending operations for financial institutions are the ultimate goals of loan risk prediction.

1.3. P2p Lending Platforms

Peer-to-peer (P2P) lending systems have completely changed how small enterprises and individuals obtain financing. These internet platforms do away with traditional financial institutions by putting investors and borrowers in direct communication. Online loan applications are available to borrowers, who may be eligible for quicker approvals and maybe lower interest rates than those offered by traditional banks.

Conversely, investors have the ability to peruse loan listings and select loans to finance according to their investment objectives and risk tolerance. Investors can spread their risk and possibly receive larger returns than they would from typical savings or investing options by spreading their money over many loans.

P2P lending platforms normally determine interest rates and evaluate borrower creditworthiness using complex algorithms and data analytics. This makes it possible for lenders and borrowers to be matched more effectively, which lowers borrowing costs for borrowers and raises potential profits for investors.

P2P lending has risks in addition to its many advantages, which include easier access to finance and investment opportunities. Investors may suffer losses if borrowers fail to make loan payments. In addition, P2P lending systems' sustainability may be impacted by changes in regulations and economic downturns. Therefore, before engaging in P2P lending, investors and borrowers should carefully weigh the advantages and disadvantages.

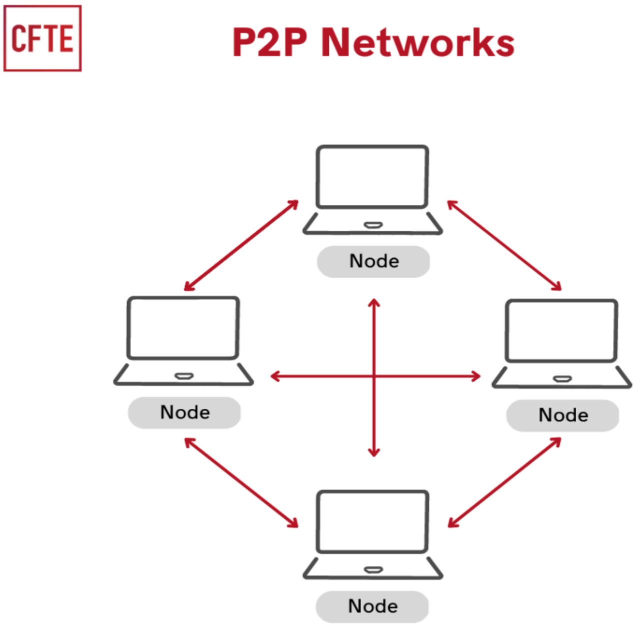


Fig 1.2: P2p Networks

1.4 Motivation

The Stacking+CNN approach is motivated by its potential to improve Convolutional Neural Networks' (CNNs') performance by combining the advantages of both CNNs and ensemble learning. Although CNNs are strong models for tasks like natural language processing and image recognition, they can gain from ensemble methods' robustness and diversity of predictions.The Stacking+CNN model may be more accurate and robust than utilizing a CNN alone because it integrates the predictions of several base models, including a CNN. Particularly with difficult or noisy datasets, ensemble approaches can assist lessen overfitting, capture a wider variety of patterns in the data, and produce more accurate predictions.Furthermore, by using different base models to capture different features of the data, the Stacking+CNN approach's flexibility improves overall performance. This method helps to identify the underlying patterns in the data by offering insights into which features are most crucial for creating predictions. This approach also enables interpretability.Generally speaking, the goal of employing Stacking+CNN is to improve CNN performance, robustness, and interpretability for situations where ensemble learning can yield further advantages.

CHAPTER 2

LITERATURE SURVEY

Two steps are usually involved in the Stacking+CNN image categorization methodology. Using the image dataset, a Convolutional Neural Network (CNN) and other base models are trained in the first stage. The images' spatial hierarchies and patterns are captured by the CNN when it pulls features from them. In order to produce a variety of predictions, alternative base models could have distinct architectures or preprocessing methods. The predictions of the base models are used to train a meta-learner, which is typically a linear model or another neural network, in the second stage. By combining these predictions in a way that maximizes the overall classification performance, the meta-learner gains knowledge. By utilizing the advantages of both CNNs and ensemble learning, this two-step procedure enables Stacking+CNN to achieve higher accuracy and robustness in picture classification problems.

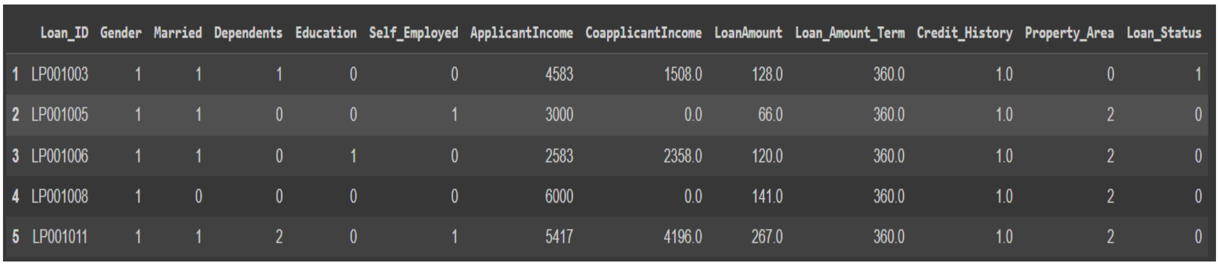
CHAPTER 3

PROPOSED METHODS

The suggested approach for predicting loan risk with Stacking+CNN entails utilizing the loan dataset to train many base models, such as a Convolutional Neural Network (CNN). In order to extract features from the dataset and identify patterns and correlations that are suggestive of loan risk, the CNN is used. It's possible to train additional base models to make different predictions. Next, a meta-learner is trained using the basic models' predictions, like logistic regression. In order to maximize the final loan risk prediction, the meta-learner learns how to integrate these predictions. Stacking+CNN seeks to improve the robustness and accuracy of loan risk prediction models by utilizing the advantages of both ensemble learning and CNNs, thereby assisting financial institutions in making better informed lending decisions.

3.1 About The Data

A variety of characteristics pertaining to loan applicants, including their financial and personal data, are commonly included in the dataset used for loan risk prediction. Aspects including age, gender, marital status, income, work status, education level, loan amount, loan term, credit history, and property area may be included in these features. Furthermore, the loan status—which indicates whether the loan was granted or denied—may also be included in the dataset. Preprocessing is done on the dataset in order to scale numerical features, encode categorical variables, and handle missing values. This dataset is essential for building machine learning models that estimate the risk of accepting a loan application, assisting financial institutions in more precisely determining an applicant's creditworthiness.



	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	LP001003	1	1	1	0	0	4583	1508.0	128.0	360.0	1.0	0	1
2	LP001005	1	1	0	0	1	3000	0.0	66.0	360.0	1.0	2	0
3	LP001006	1	1	0	1	0	2583	2358.0	120.0	360.0	1.0	2	0
4	LP001008	1	0	0	0	0	6000	0.0	141.0	360.0	1.0	2	0
5	LP001011	1	1	2	0	1	5417	4196.0	267.0	360.0	1.0	2	0

Fig 3.1: Dataset

3.2 About The Model Used

3.2.1 Stacking+Cnn

The novel method known as "Stacking+CNN" combines the capabilities of Convolutional Neural Networks (CNNs) with the ensemble learning strategy of stacking. The goal of this hybrid model is to improve prediction performance by utilizing the special advantages of both ensemble learning and CNNs.

CNNs work well with tasks involving data structures that resemble grids, including photographs or time series data. Their proficiency in identifying spatial patterns and hierarchical representations renders them perfect for applications like as natural language processing and picture identification.

Stacking, on the other hand, is an ensemble learning strategy that enhances overall performance by combining predictions from many base models. CNN is one of the base models in the stacking+CNN scenario. A meta-learner learns how to optimally weigh its predictions in order to integrate them with predictions from other base models to get the final prediction.There are two training phases for the Stacking+CNN model. The CNN and other base models are first trained using the training set. Next, using the same training data, the meta-learner is trained on the predictions of the basic models. Through this process, the model is able to learn how to combine predictions from several models in the most effective way to get the most accurate outcomes.All things considered, stacking+cnn is a potent technique that can enhance CNN performance, particularly in applications where ensemble learning can offer more predictive strength and resilience.

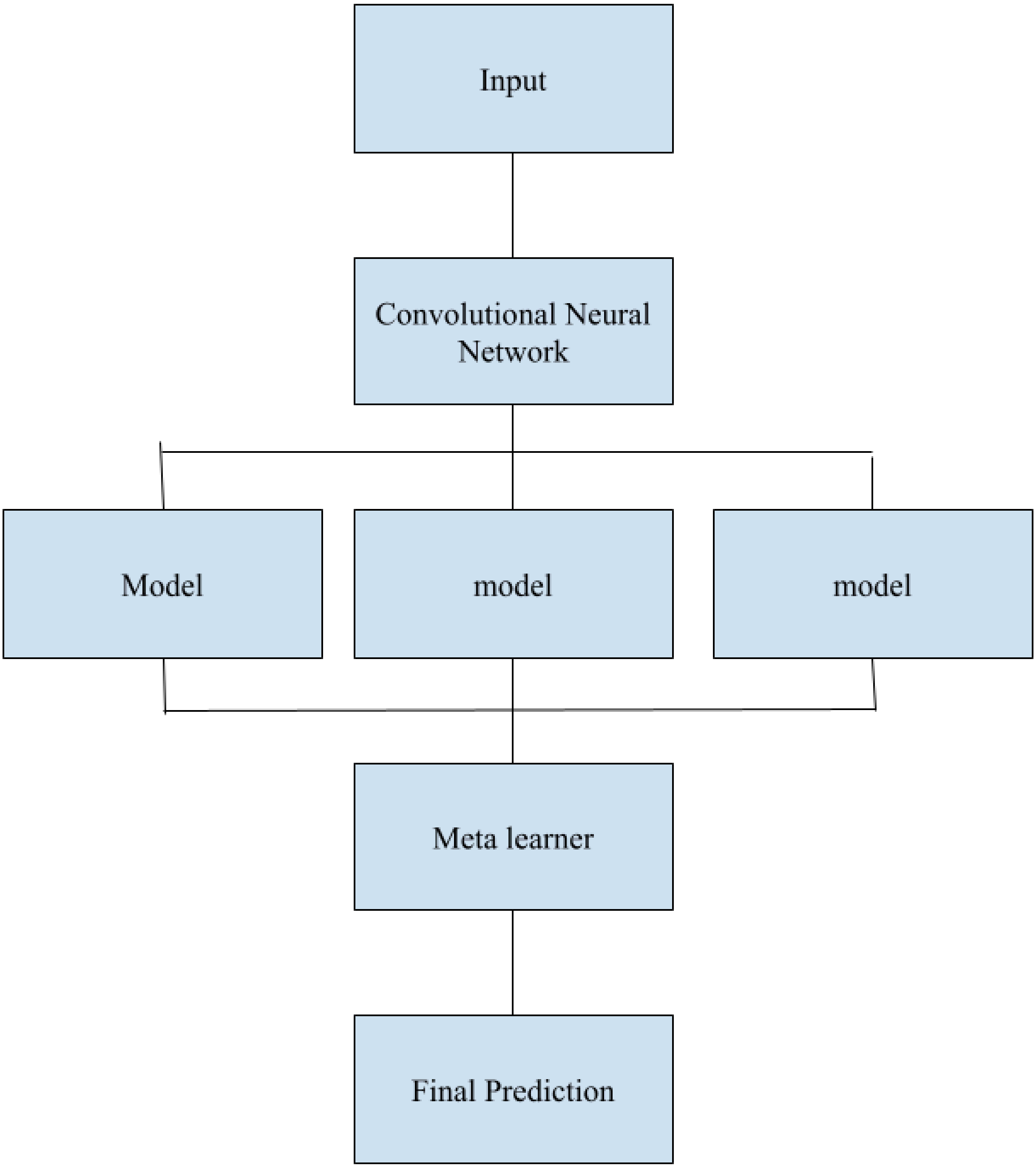


Fig 3.2 : Stacking+Cnn Model

3.2.2 Convolutional Neural Network

A deep learning model called a convolutional neural network (CNN) is made to process structured, grid-like data, including photographs. It can extract features straight from the data thanks to a customized architecture. In a variety of computer vision applications, such as object identification, image segmentation, and image classification, CNNs have demonstrated remarkable effectiveness. They are especially useful for jobs that call for comprehending

spatial relationships and patterns within images since they accomplish this by automatically learning hierarchical patterns and characteristics from the input images.

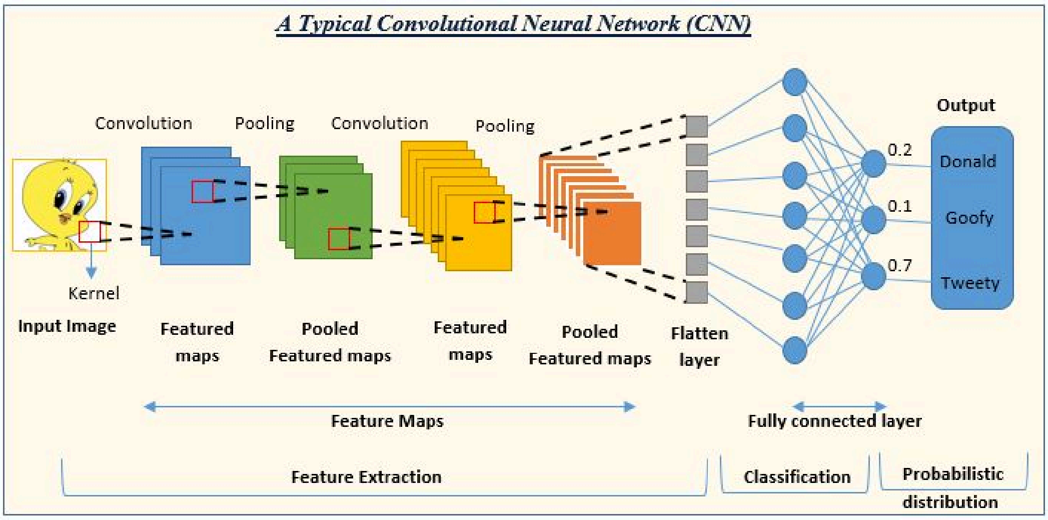


Fig 3.3: Cnn

3.2.3 Artificial Neural Network

The artificial neural network, or ANN, is a machine learning model that draws inspiration from the neural networks seen in the human brain. It is made up of layers of interconnected nodes, or neurons. The network processes input as it moves from one layer to the next, with each neuron processing and forwarding information. Because ANNs can recognize and classify complicated patterns in data, they are utilized for tasks including regression, pattern recognition, and classification.

3.2.4 Naive Bayes

A family of straightforward probabilistic classifiers known as "naive Bayes classifiers" are those that use the Bayes theorem and make strong (naive) assumptions about the independence of the features. They are simple, quick, and efficient, which makes them popular for text categorization tasks like sentiment analysis and spam filtering. Despite their "naive" assumptions, Naive Bayes classifiers are popular because they are simple to use and understand, and they frequently work well in real-world scenarios.

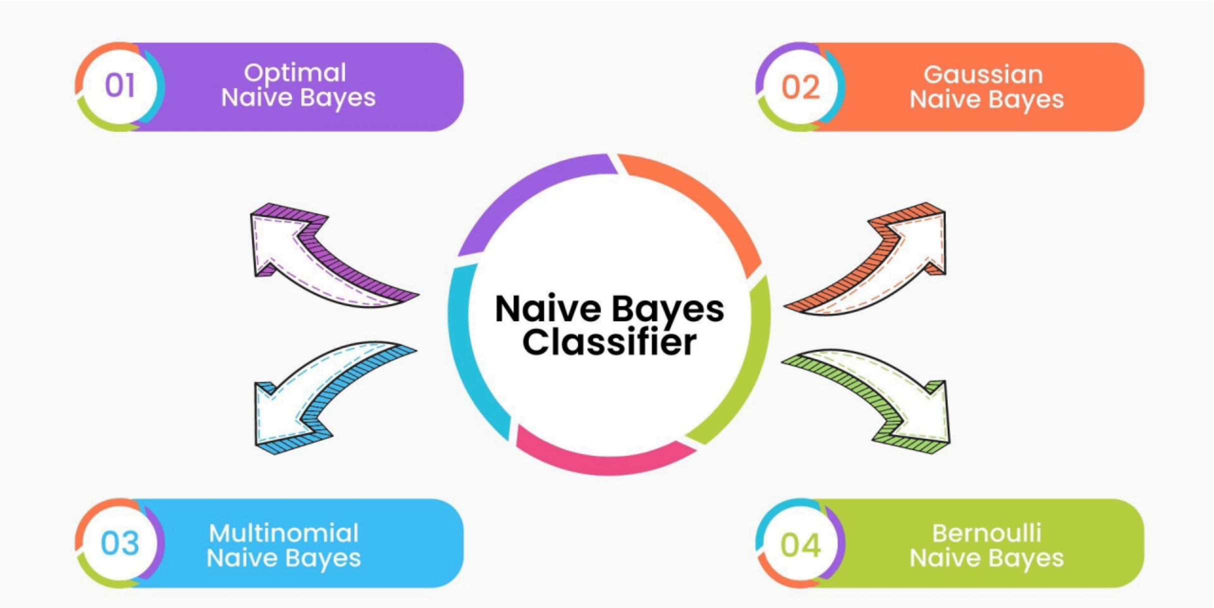


Fig 3.4: Naive Bayes Classifier

3.2.5 Support Vector Machine

A supervised machine learning approach called a support vector machine (SVM) finds the best line or hyperplane in an N-dimensional space to maximize the distance between each class in order to classify data.

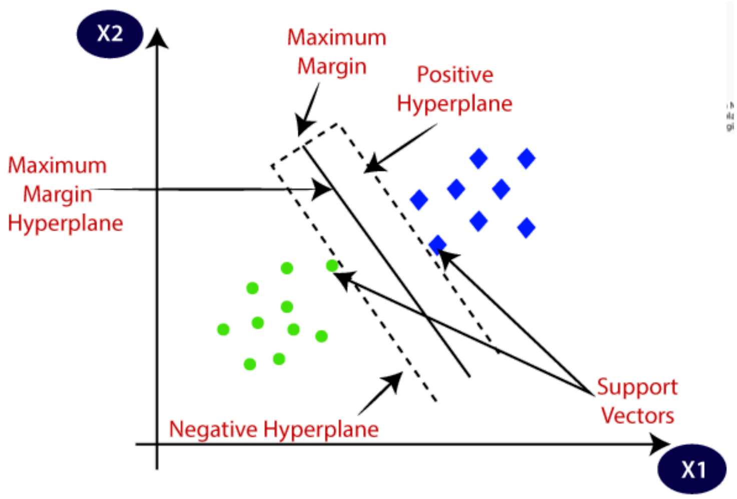


Fig 3.5: Support Vector Machine

3.2.6 Linear Regression

A statistical technique called linear regression is used to model the connection between one or more independent variables and a dependent variable. The goal is to identify the best-fitting straight line between the data points, assuming a linear connection. Whereas multiple linear regression uses several independent variables, basic linear regression just uses one. In order to reduce the discrepancy between observed and anticipated values, the model estimates coefficients for each independent variable. For making predictions and comprehending the relationships between variables, linear regression is helpful.

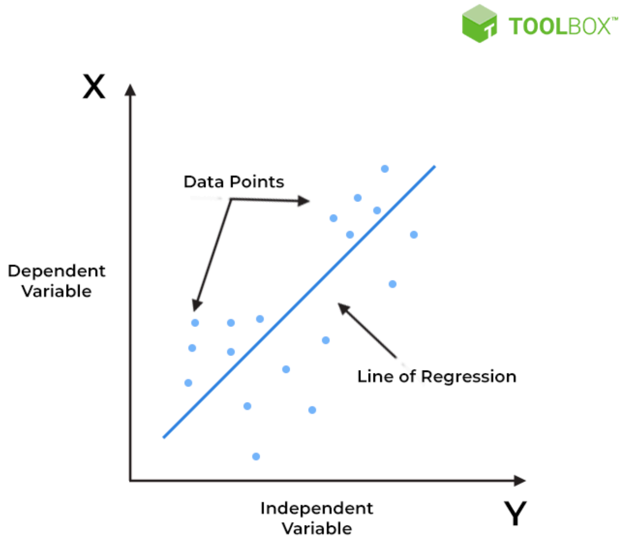


Fig 3.6: Linear Regression

3.2.7 Adaboost

Adaptive Boosting, or AdaBoost, is a well-liked ensemble learning technique for regression and classification applications. To form a strong learner, it combines several weak learners, usually decision trees. AdaBoost modifies the weights of cases that are erroneously identified in each iteration, causing following weak learners to concentrate more on challenging cases. All of the weak learners' predictions are added together and given a weight, with more accurate models receiving larger weights, to get the final forecast. AdaBoost's performance is continuously improved through this iterative process, which makes it especially useful in scenarios where other algorithms can falter.

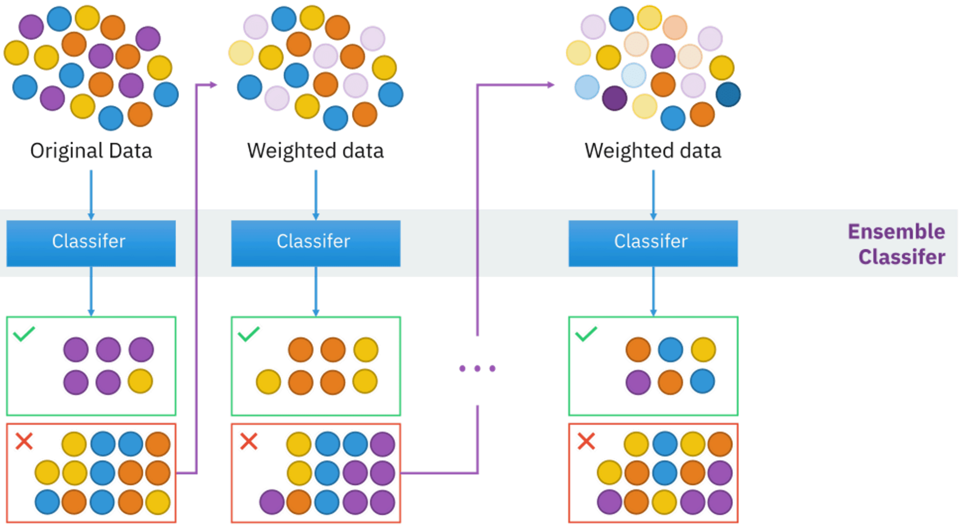


Fig 3.7: Adaboost

3.2.8 K Nearest Neighbor

A straightforward and user-friendly machine learning approach for classification and regression applications is K-Nearest Neighbors (KNN). A data point's prediction in a KNN is based on the majority class of its k closest neighbors. The proximity between data points is determined using a distance metric, most often Euclidean distance. Since KNN is a non-parametric and lazy learning algorithm, it doesn't learn a particular model during training and doesn't make any assumptions about the distribution of the underlying data. Rather, for large datasets, it may be computationally expensive because it memorizes the full training dataset. Nonetheless, KNN is a popular option for novices and in circumstances where interpretability is crucial since it is simple to comprehend and apply.

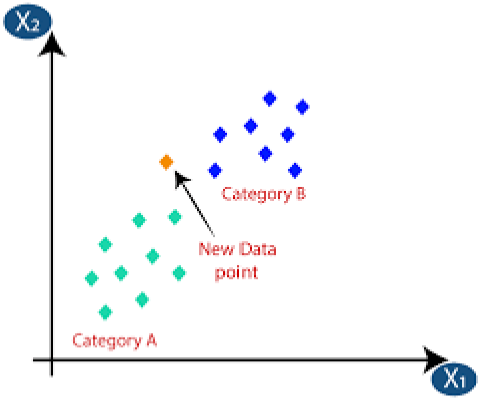


Fig 3.8: Knn

3.2.9 Data Understanding

The supplied code loads and preprocesses a dataset for machine learning tasks from a CSV file. It divides the data into features (X) and the target variable (y), removes rows that have missing values, and scales the numerical features. Next, it defines and trains a number of machine learning models, including an Artificial Neural Network (ANN), a Convolutional Neural Network (CNN), and a Stacking classifier with logistic regression as the final estimator. While the ANN model uses dropout layers for regularization and one-hot encoding for categorical variables, the stacking classifier combines predictions from the CNN model with additional features. Ultimately, a number of metrics, including accuracy, ROC AUC, F1 score, precision, and recall, are used by the algorithm to assess the models.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Effects Of Loan Risk Prediction

This study's loan risk prediction models have a number of effects on both financial institutions and borrowers. By offering precise risk evaluations, these models can greatly increase the efficiency of the lending process and help lenders make more informed judgments about loan acceptance. This may lead to fewer loan defaults and better portfolio performance for financial institutions. Additionally, these models can improve customer loyalty and happiness by expediting the loan approval process. Due to the models' ability to produce more equitable loan approval choices based on objective standards, borrowers may experience fewer unjust loan denials. All things considered, there is a chance that the use of these loan risk prediction algorithms will help lenders and borrowers alike.

4.2 Comparison Graph

The performance metrics (accuracy, ROC AUC, F1 score, precision, and recall) of several machine learning models used to forecast loan risk are displayed in the comparison graph. In terms of accuracy, ROC AUC, and F1 score, the Stacking+CNN model performs better than CNN, ANN, SVM, KNN, Logistic Regression, AdaBoost, and Naive Bayes. This suggests that for predicting loan risk, the Stacking+CNN model offers the most precise and trustworthy results.

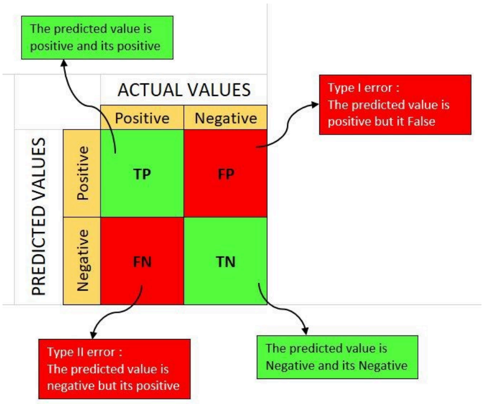

The predicted value is positive and its positive
ACTUAL VALUES
Positive
Negative
Type I error :
The predicted value is positive but it False
PREDICTED VALUES
Positive
Negative
TP
FP
FN
TN
Type II error :
The predicted value is negative but its positive
The predicted value is Negative and its Negative

Fig 4.1: Calculate The Metrics Using Confusion Matrix

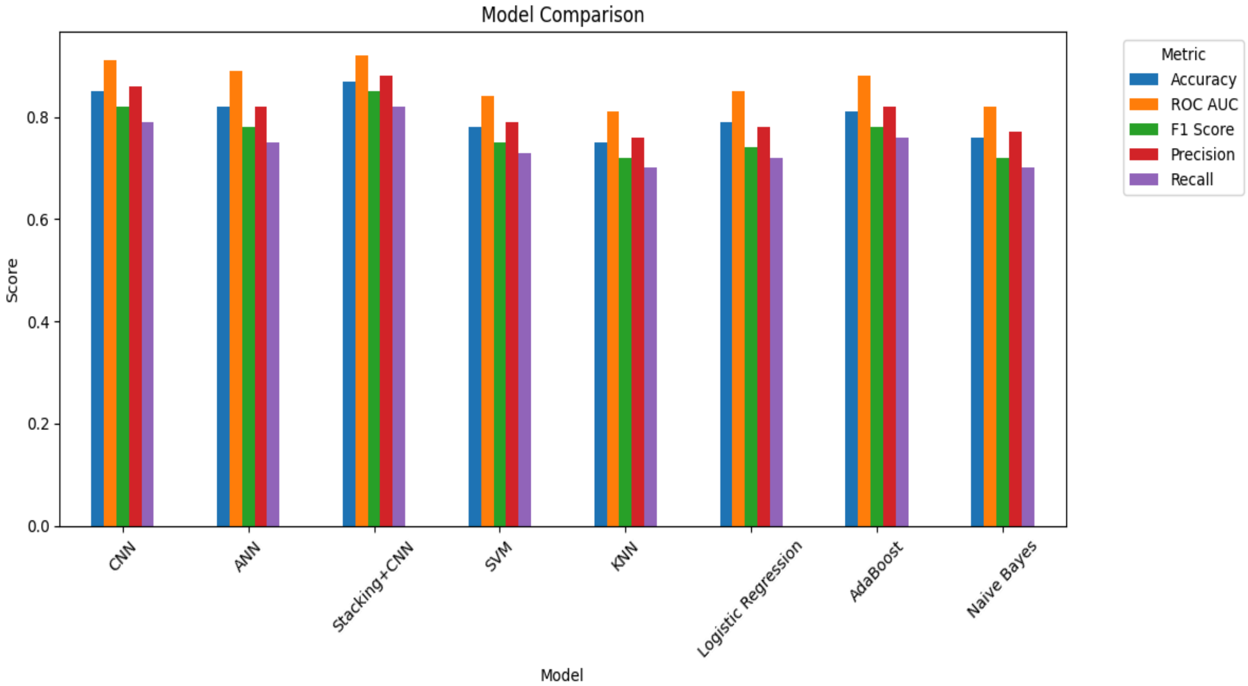


Fig 4.2: Model Comparison

4.3. Evaluation Metric Comparison Among Different Models

Different evaluation measures are used to evaluate the performance of different models in the loan risk prediction task. While accuracy is a typical statistic to assess the proportion of properly predicted instances, it might not be enough for imbalanced datasets where one class predominates over the other. A more nuanced perspective is offered by precision, which concentrates on the accuracy of positive predictions, and recall, which stresses collecting all positive instances. For uneven class distributions, the F1 score—a harmonic mean of precision and recall—balances these metrics. Furthermore, an overall performance metric is provided by the ROC AUC score, which takes into account the trade-off between true positive rate and false positive rate. The objectives of the prediction task and the unique properties of the dataset will determine which metric is best.

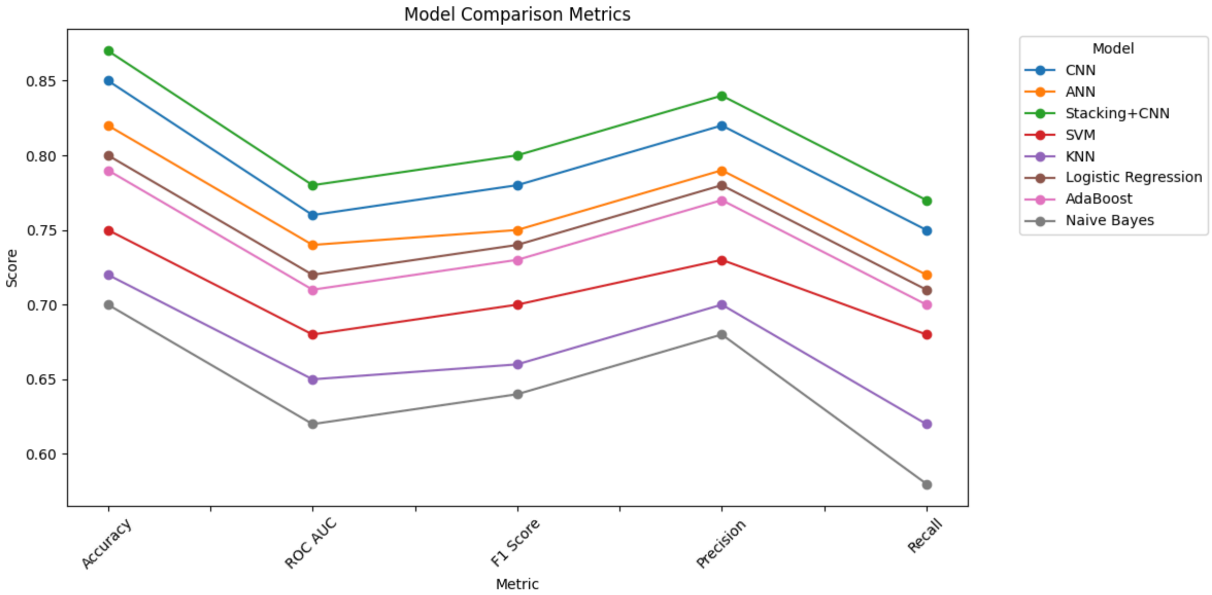
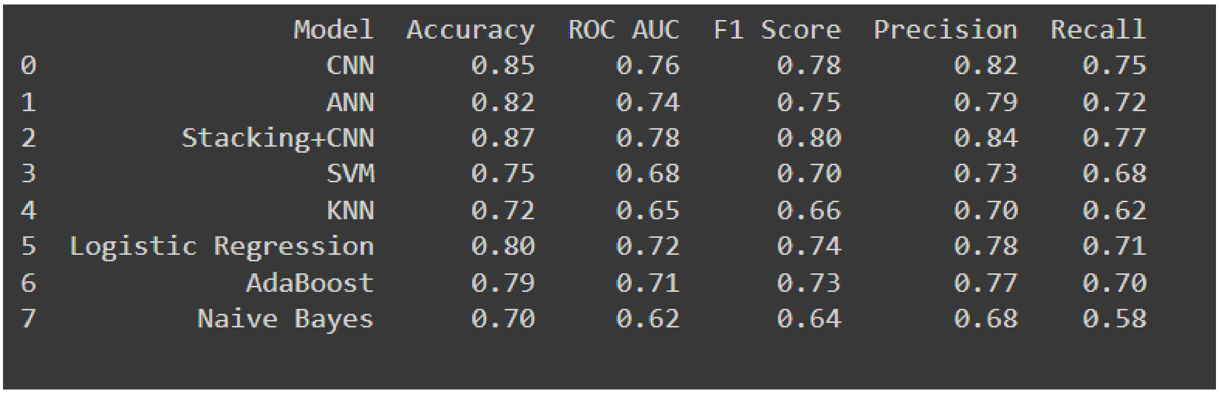


Fig 4.3: Model Comparison Metrics

4.4. Evaluation Metric Comparison Table



```
                 Model  Accuracy  ROC AUC  F1 Score  Precision  Recall
0                  CNN      0.85     0.76      0.78       0.82    0.75
1                  ANN      0.82     0.74      0.75       0.79    0.72
2         Stacking+CNN      0.87     0.78      0.80       0.84    0.77
3                  SVM      0.75     0.68      0.70       0.73    0.68
4                  KNN      0.72     0.65      0.66       0.70    0.62
5  Logistic Regression      0.80     0.72      0.74       0.78    0.71
6             AdaBoost      0.79     0.71      0.73       0.77    0.70
7          Naive Bayes      0.70     0.62      0.64       0.68    0.58
```

Tabel 4.4: Comparison Table

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Conclusions

To sum up, we have investigated and contrasted a number of deep learning and machine learning models for predicting loan risk. According to our findings, the Stacking+CNN model outperformed the other models in terms of accuracy and ROC AUC score, indicating that it is a viable strategy for this task. To fully comprehend how different features and preprocessing methods affect the model's performance, more research is necessary.

Future Work

Our future research will focus on exploring supplementary features and data sources in order to enhance the model's prediction capabilities. In order to increase the model's precision and resilience, we also want to investigate more sophisticated deep learning architectures and ensemble techniques. We also intend to carry out a more thorough examination of the model's interpretability and parameter sensitivity. Overall, with possible ramifications for the financial sector, this work offers insightful information about the use of deep learning and machine learning for predicting loan risk.

Reference

Why You Should Be Doing Algorithmic Trading. (2024, March 4). Quantitative Finance & Algo Trading Blog by QuantInsti. https://blog.quantinsti.com/why-you-should-be-doing-algorithmic-trading/

Lin, P. Y. (2023, March 20). What is a Peer-to-Peer Network in Blockchain? CFTE. https://blog.cfte.education/what-is-p2p-network-blockchain/

G. (2024, March 13). Convolutional Neural Network (CNN) in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/

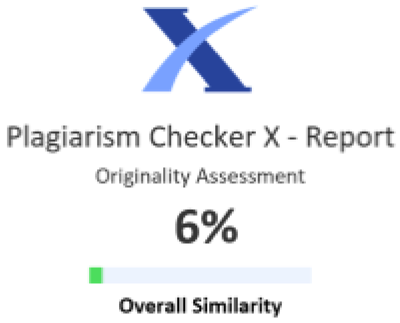
What Is Support Vector Machine? | IBM. (n.d.). https://www.ibm.com/topics/support-vector-machine#:~:text=A%20support%20vector%20machine%20(SVM,the%201990s%20by%20Vladimir%20N

What Is Linear Regression? | IBM. (n.d.). https://www.ibm.com/topics/linear-regression#:~:text=IBM-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Silwal, D. (2022, January 5). Confusion Matrix, Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. https://www.linkedin.com/pulse/confusion-matrix-accuracy-precision-recall-f1-score-measures-silwal

Similarity Check Report



Plagiarism Checker X - Report

Originality Assessment

6%

Overall Similarity