A Comparative Analysis for Predicting Loan Default Risks using Machine Learning Algorithms

1st Ashish Garg
Computer Science and Engineering
Graphic Era Deemed to-be University
Dehradun, Uttarakhand, India
geuashishgarg@gmail.com

2nd Rameshwari Kapoor
Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand, India
rameshwari.kapoor09@gmail.com

3rd Nilesh Bhanot
Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand, India
nileshbhanot18@gmail.com

4th Ayush Maheshwari
Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand, India
ayushmaheshwari07062002@gmail.com

***Abstract*—The process of loaning credit by banks helps individuals and business to grow financially and make large payments and investments upfront that they might not have been able to make by themselves. This helps the economy by maintaining cash flows during the lean periods. However, crediting a loan comes with a risk that the client or the business might not be able to repay the loan in the future leading to loan defaults. To mitigate this issue, banks for a long time have conducted background financial checks on clients manually. In 21st century, with the availability of surplus data and resources, machine learning and deep learning models can be trained to perform the same task hence, reducing time and effort exponentially than manual work to make the decision whether to disburse the loan or not. Finance sector in India comprises mainly of commercial banks that generate most of their business through the interest received on loans disbursed to people and organisations. However, credit lending comes with risk of consumer defaulting the loan i.e., inability to pay back the due sum of money agreed upon by both the parties. To ensure that the borrower will repay the loan on the proposed terms is the biggest challenge faced by banks in today's ever-changing world. This paper puts forward an approach to leverage different machine learning (ML) algorithms to predict which consumers are likelier to default their loans using historical financial data i.e., their credit and debit details, bank transactions, previous credit history, current bank loans and income of the client. This study makes use of the Credit Risk Model Stability data available on Kaggle provided by the Home Credit finance provider, founded in 1997. The research is a comparative analysis between different types of algorithms in machine learning. Results of this study can be scaled and applied to a real-world dataset and holds the immense potential to revolutionise the financial industry.**

***Index Terms*—credit, loan default, gradient-boosting, bayesian-learning, machine-learning**

I. Introduction

The health of an economy is dependent upon its financial sector. A weak financial sector results in a declining economy and the stronger the finance sector, safer is the economy of a nation. A healthy and stable economy requires a strong financial sector. The finance sector strengthens itself by providing loans to business owners, mortgages to homeowners and by issuing insurance policies to the population and against assets. These services provided by the finance sector forms the backbone of the economy. India's finance sector is dominated by commercial banks which accounts for more than 64% of the sector's total assets. Commercial banks are the main players in Indian financial sector. Most of their revenue is earned from the interest on loans that is extended to individuals and corporates. However, granting loans has its own pros and cons. One of the cons include default risk. Default risk means when someone takes a loan but due to some reason is unable to pay back the agreed amount. To minimize this risk and ensure that the loan is payed-off according to the agreed terms and conditions, gave us the motivation to work on this project. To address the challenge, this paper leverages a combination of eight different machine learning and deep learning algorithms to predict the likelihood of loan default. The algorithms used in this study are Deep learning models like Long Short-term memory, Lazy Learning algorithms, gradient boosting models as well as Decision Trees and Random Forest. The prowess of machine learning models in this ever-changing financial sector can be attributed to two major reasons. The first reason is, with the advancement of technology and increasing online transactions, banks are now able collect more data than ever from internal and external data sources which can be easily processed by teaching a computer to do it. The second reason is the success of ML models in similar applications like stock price prediction and credit card fraud detection in the banking sector. This explorative research uses real-world banking data provided by the Home Credit, an international consumer finance provider on Kaggle. The dataset consists of masked data of actual clients split into 32 training and 36 testing csv files. For each client id, there exists a dependent target class having values, 0 (client repays the loan) and 1 (client defaults the loan) that is to be predicted. After collecting the required data, we apply data mining, preprocessing and feature engineering methods to create a dataset of filtered features be fed to a machine learning model. We compare

the efficiency and performance of the 8 models, out of whom, gradient boosting machines achieve nearly 99% accuracy, thus showcasing the promise of scaling these models to a real-world use case in the finance sector.

II. Literature Review

Asha RB et al. (2021) compared the efficiency of three algorithms namely Support Vector Machines accompanied by K-Nearest Neighbours and Artificial Neural Network. The experiment resulted in ANN performing the best with an accuracy of 99% closely followed by KNN and SVM. Ebenezer Esenogho et al. (2022) used imbalanced dataset having information of European Credit Card Clients. They potrayed the usage of Multi-layer Perceptron, Support Vector Machines, a variety of Long short-term memory models along with AdaBoost and Decision Tree. Out of all these, Proposed LSTM Ensemble model outperformed all others with an accuracy score of 99%. Lin Zhu et al. (2019) evaluated multiple models that include Random Forest as well as Logistic Regression, SVM and Decision Tree. The outcome of this study showcased the prowess of Random Forest model over the other aforementioned models with an accuracy of 98%. John O. Awoyemi et al. (2017) used hybrid sampling to handle imbalanced data. Three algorithms named Naïve Bayes, KNN and Logistic Regression have been compared. Their experimentation revealed the success of K-Nearest Neighbours model with an accuracy of 97.9% over Naive Bayes. Jing Gao et al. (2021) deployed XGBoost and LSTM models during their comparative study. The outcome shows that XGBoost-LSTM model predicts credit card default with an accuracy of 95.4%, whereas XGBoost alone predicts with an accuracy of 89.5%. V.A. Kandappan et al. (2021) made use of Bidirectional LSTM to predict loan defaults. The conclusion of that study reveals that LSTM achieved a promising accuracy of 94%.

Anushi Jain et al. (2022) evaluated the effectiveness of five machine learning algorithms namely, Artificial Neural Networks, alongwith Logistic Regression, gradient boosting and support vector machines. Their study reveals that Logistic Regression is the best model to predict Loan default with an accuracy of 88.89% followed by Random Forest and XG Boost. Md. Golam Kibria et al. (2021) implemented deep learning models alongside two different models i.e, Logistic Regression and Support Vector Machines. The outcome of this experiment dictates that deep learning models performed superior to that of machine learning models showcasing an accuracy of 87.10% over 86.23%. Huannan Zhang et al. (2020) illustrated the use of three algorithms i.e., Decision Trees together with Random Forest and also Logistic Regression. The conclusion of the study exhibits the efficiency of the Random Forest to predict loan default with an accuracy of 86% considerably exceeding the performance of the other two models. Bhoomi Patel et al. (2020) analysed the proficiency of four models to predict the probability of a loan default by consumers. Among the deployed models, CatBoost Classifier outperformed other algorithms. It potrayed an accuracy of 84.045% beating the other algorithms.

Yanash Azwin Mohmad (2022) leveraged Long Short Term Memory model with an addition of three other algorithms as well: Multi-layer Perceptron along with Support Vector Machine and Random Forest. Out of the models compared, LSTM model showed the promise with an accuracy of 82.4% in forecasting of late payments of loans. Abhishek Shivanna et al. (2020) used different algorithms including Deep Support Vector Machine (DSVM) to achieve an accuracy of 82.2% in predicting the defaulters. They also deployed Boosted Decision Trees, along with Averaged Perceptron and Bayes Ponit Machine algorithm techniques. The results shows that out all the models DSVM can best predict defaulters with an accuracy of 82.20%.

Yue Yu (2020) compared and contrasted the features of four algorithms named Logistic Regression as well as Random Forest in company with Decision Trees and AdaBoost. The yield of this experimental study portrays the prowess of Random Forest classifier portraying an accuracy of 82.12%. Saurabh Arora et al. (2022) evaluated and examined six machine learning algorithms. The conclusion revealed the proficiency of Support Vector Machine model to predict credit card default with an accuracy of 82% closely followed by Logistic Regression and Random Forest models with respective accuracies of 81% and 80%. Theoneste Ndayisenga (2021) portrayed the usage of machine learning algorithms such as lazy learning algorithms, gradient boosting algorithms and ensemble learning models in addition to Support Vector Machines and simultaneously, Logistic Regression as well. It can be concluded from the study that Gradient Boosting algorithms perform the best in predicting bank loan defaults. Mehul Madaan et al. (2021) studied the functioning of Decision trees and random forest models in predicting bank loan defaults. The findings of their experimentation suggests that Random Forest outclassed Decision Trees with an accuracy of 80% over 73% respectively. Malik Mubasher Hassan et al. (2020) examined the usage of Artificial Neural Networks to predict customer defaults. The result showed that ANN can predict the customer default with an accuracy of 77.9%.

Alžbeta Bačová and František Babič (2021) have used Random Forest, AdaBoost and XGBoost for predictive analysis for credit card default. The results illustrated similar performance between all the models used. Lili Lai (2020) has compared AdaBoost, XGBoost, Random Forest, KNN and Multi-Layer Perceptron algorithms to predict loan default. The conclusion is that AdaBoost outperformed all the other algorithms followed by XGBoost. Luca Barbaglia et al. (2021) used Penalized Logistic Regression, Gradient Tree Boosting and XGBoost for a highly unbalanced dataset of 12 million residential mortgages. The result revealed that XGBoost and Gradient tree Boosting outperformed Penalized Logistic Regression model. Hyeongjun Kim et al. (2020) performed statistical analysis on machine learning models in addition to Binary Response models and also performed Discriminant analysis and also employed hazard models. Their team employed the algorithms, namely, Decision trees along with Support Vector Machines or SVM. The performance of Artificial Neural Network is also

represented in their research.

Abhishek Agarwal et al. (2020) experimented five machine learning models during the course of their study, namely, Logistic Regression and Naive Bayes as well as Decision Trees in addition to Random Forest and KNN. This suggested the changes in model performance after the application of Principal Component Analysis on the attributes of the dataset. The study concluded that Logistic Regression outclassed Decision Tree in both the scenarios. Mohammad Ahmad Sheikh et al. (2020) showcased the deployment of Principal Component Analysis to analyse the importance of feature engineering on the efficieny of a model. The conclusion of is that the model is marginally better after applying PCA.

Basna Mohammed Salih Hasan et al. (2021) gave the basic idea of Principal Component Analysis and described some of the related concepts. Farzana Anowar et al. (2021) performed a conceptual comparison between eleven different dimensionality reduction algorithms, PCA being one of them. Robert Reris et al. (2015) discussed the problems faced regarding optimisation during PCA stage. They employed multiple geometric angles to the problem in their research.

Machine Learning algorithms are used in various other fields. One such application is to classify skin diseases as specified by Bhadula* et al. (2019). They portrayed their work on the following models, Convolutional Neural Networks and other similar algorithms such as Logisitic Regression and Naive Bayes alongwith Random Forest and Support Vector Machines. M.Kamal et al. (2022) in their experimentational study discussed the deployment of ML algorithms in the field of healthcare. They developed models including Random Forest classifier and Support Vector classifier in identifying the conditions for classification of Alzheimer's disease. D. Bordoloi et al. (2022) recognized the emergence of ML and DL algorithms in the field of healthcare systems. They reviewed the advancements in ML and DL for achieving multi-objective goals in the field of healthcare.

III. Methodology

In the past few years, Machine learning algorithms have revolutionized the finance sector by offering financial institutes a powerful data-driven tool to help with the enormous tasks of predicting loan defaults by a client. The workflow of the applied methodology is depicted in the flowchart in Fig 1.

A. Data Collection

The dataset used in the model is the Home Credit – Credit Risk Model Stability provided by the Home Credit organisation on Kaggle. It consists of 32 training files and 36 testing files. All the files are available in csv and parquet format. The files consist of data collected for a particular case id from two types of sources: internal and external data sources. The training files are used for training the models. These are further divided into two parts in the ratio 75:25 for training and testing purposes respectively.

B. Data Mining

The csv files extracted from the dataset are needed to be merged and converted into a format that can fed into the machine learning models. To merge all the files together, a "train_base.csv" file is provided that contains the target variable. The files were merged into a single dataset using the python library, polars. Polars is a python library written in Rust and uses a multi-threaded query engine for fast and effective parallel execution.

C. Preprocessing

After converting all the csv files into a single polars dataframe, the data consists of null values, categorical string data and dates. This data needs to be pre-processed and converted into a usable format. Firstly, the dataset is transformed back into pandas dataframe for label encoding the columns with string data. Label Encoding is a preprocessing technique in which string data is converted to numerical data by allocating a unique index value to each unique string value in a column. Another data preprocessing technique known as standard scaling is used. Standard scaling is also known as Z-score normalisation. The columns of the dataset with numerical values are scaled such that they are normalised and all the values in the dataset have the same scale of standard deviation equal to 1 and an average or mean value equal to 0. The formula for standard scaling is:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where, x = original value of the feature,
μ = mean of the feature values, and
σ = standard deviation of the feature values
Standard scaling is used to normalise distribution of values, reduces dominance of a single feature or multiple features in the dataset and improves converge rate of the model.

D. Feature Engineering

The final dataset after cleaning and preprocessing consists of 303 independent features, 1 target variable and over 15 lakh records. The size of the dataset was reduced from 15 lakh records to 1 lakh records so that dataset is balanced and there is equal representation between the two classes. Faster convergence and improved training times can be achieved using a dimensionality reduction algorithm known as Principal Component Analysis or PCA. PCA is a method which helps in reducing the number of columns or features in a dataset while maintaining all or majority of critical information. The benefits of performing PCA include faster model training times since dataset size is reduced, hence, the final trained model is of small size as it has less parameters during training. PCA can be lossy or lossless in nature depending on the dataset. Using PCA, top 100 features were extracted to train the machine learning models, some of which are given in the table 1.

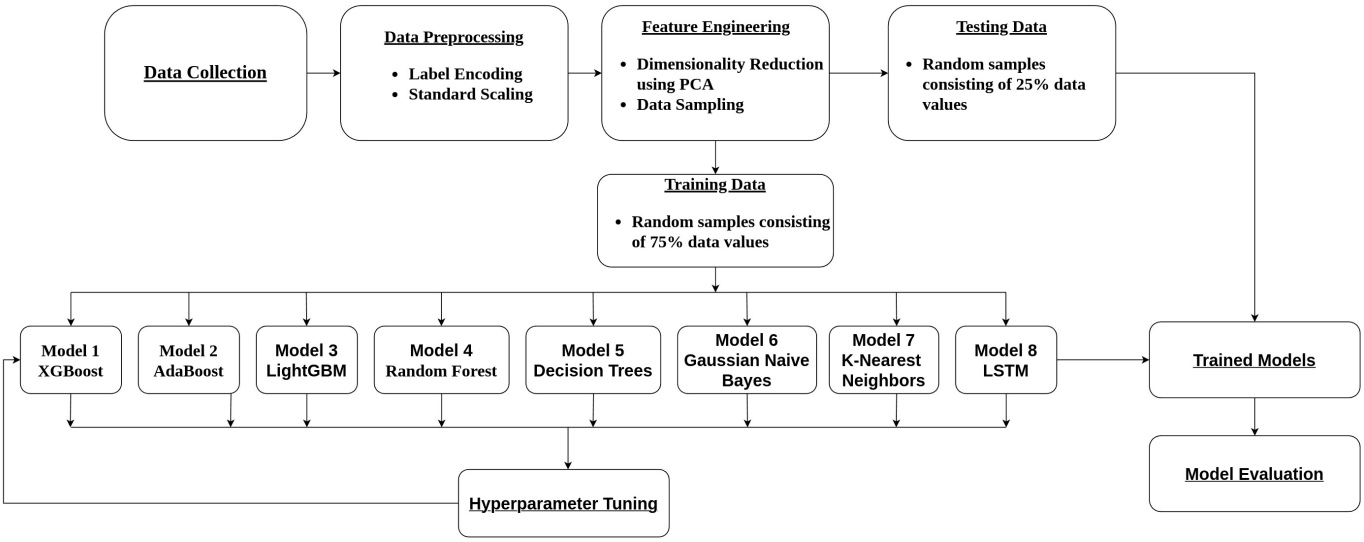

Fig. 1. Workflow of the Proposed Methodology.

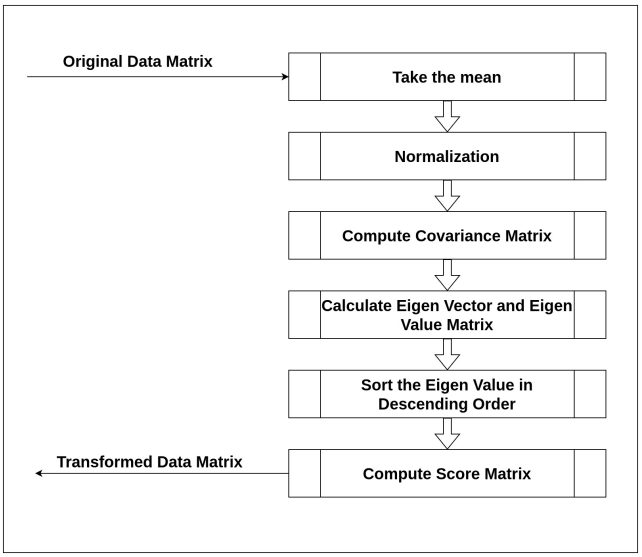

Fig. 2. Principal Component Analysis.

TABLE I
SOME IMPORTANT FEATURES OF THE DATASET

Feature Name	Feature Description
credamount_770A	Loan amount or credit card limit
cntpmts24_3658933L	Number of months with any incoming payment in last 24 months
bankruptcy_history	Bankruptcy history of the client
age	Age of the client
education_level	Education level of the client
credit_score	Credit score of the client
employment_history	Employment history of the client
debt_to_income_ratio	Debt-to-income ratio of the client
income	Income of the client
payment_history	Payment history of the client
late_payment_history	Late payment history of the client

IV. PROPOSED MODEL

Considering the features and complexity of the dataset, the models deployed and showcased in this experimental study for the purpose of comparative analysis: are ensemble learning, gradient boosting, Bayesian learning, lazy learning and deep learning algorithms. The dataset has been segregated into two segments, a set of records consisting of training data which contains labelled values for training the model and another set of records consisting of testing data that contains unlabelled values for testing the trained model. This split is done in the in the ratio of 75:25 respectively such that model can better generalise the features of the dataset.

A. Decision Trees

A Decision tree is a hierarchical data structure in the form of a tree that is used in real-world scenarios interchangeably, for classification and regression problems. All the internal nodes of this tree-like structure depicts a decision which is taken on the basis of a feature of the dataset and all the leaf nodes represent the output of the predictor or classifier model. It splits the dataset features recursively into subsets based on the feature that best divides the data i.e., providing the maximum information gain into separate classes. The dataset is divided at each step such that it maximises the information gain.

$$E(S) = -p_{(+)} \log p_{(+)} - p_{(-)} \log p_{(-)} \tag{2}$$

$$\text{Information Gain} = E(Y) - E(Y \mid X) \tag{3}$$

B. Random Forest

Random forest machine learning model is a mathematical algorithm that is based on concept of ensemble learning which creates various weak learner decision tree models and depending on the outputs of weak learner models, the output of the complete model is generated. The output generated by a random forest model in the case of a classification problem is dependent on the voting by the multiple decision trees and for a regression task, the output is the mean of the output of the internal decision trees. It is indifferent to noisy data and generalises the model reducing over-fitting.

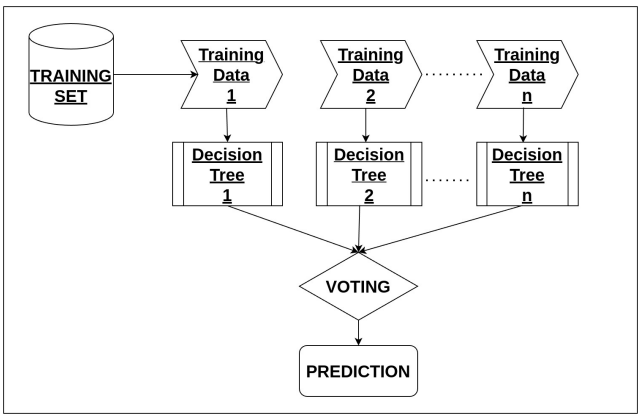

Fig. 3. Random Forest Model.

C. Gradient Boosting

Gradient boosting is a boosting algorithm in which a strong model is developed by sequential learning of weak learning models. It combines weak learner models and optimises them to minimise the value of a loss function. The sub-classes of this algorithm used in this study are: Xtreme Gradient Boosting, Ada Boost and LightGBM. The aim of these algorithms is to minimise the loss function, in this case, for binary classification is log loss function. The log loss function heavily penalises the wrong classifications.

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i) \cdot \log(1-p(y_i)) \qquad (4)$$

D. LightGBM

The lightgbm is a gradient-boosting framework based machine learning model used to classify data. It is well suited for tasks that make use of large datasets. This classifier is majorly deployed due to its efficiency, speed and highly accurate results. In this use case, the gradient boosting mechanism of lgbm minimises the log loss function and employs an EarlyStopping method to prevent overfitting by the model. It also shows one of the fastest training times out of the 8 ML models deployed.

E. AdaBoost

AdaBoost is an algorithm which stands for Adaptive Boosting and it was first developed by two developers, namely, Yoav Freund and Robert Schapire. AdaBoosting is an ensemble learning based machine learning framework that uses predictions from weak learner models to form a strong learner model. The wrong predictions are used to consecutively change the weights of the model in order to improve its accuracy. In this use case, we have employed 50 weak learner estimators to predict whether the given client will default their loan or not. The AdaBoost model provides the highest accuracy achieved in this study among all the other models.

F. XGBoost

XGBoost is an abbreviation for eXtreme Gradient Boosting. It is a highly efficient gradient boosting framework that widely used for its speed and performance. It provides high speed and accuracy and is scalable to large datasets because it supports parallel computing. It uses L1 and L2 regularisation that helps prevent the model to overfit the given data. In the given use case, it minimises the log loss function shown in Fig. 6 due to the nature of the problem, i.e., Binary classification. This machine learning model provides fastest training time and is also among the best performing models in this study.

G. Gaussian Naive Bayes

Gaussian Naive Bayes is an algorithm that functions on the principle of Bayes' theorem of probability and is used for probabilistic modelling. It is used for performing classification tasks effectively. It takes into consideration that the input features have their independent characteristics and follow a Gaussian (normal) distribution curve. It works by calculating the probability for each class for a set of input features and returns the class with maximum probability.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)} \qquad (5)$$

H. K-Nearest Neighbours

K-Nearest Neighbours is a lazy learning model that allocates a category to a set of values in the dataset based on its calculation of distance metric of that data point from all other points and then evaluating the majority of its k-nearest neighbours. The distance metric can be of different types: Euclidean distance, Manhattan distance and Chebyshev distance. It is a simple and easy to implement algorithm. The distance metric used in this research is Euclidean distance. It does not store the dataset in memory and is simple to implement.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (6)$$

I. Long Short-Term Memory

Long Short-Term Memory is an altered version of Recurrent Neural Network architecture that helps to encapsulate the context in sequential and long-term data. It consists of four types of gates or cells to retain long-term information and process long sequential data with ease, namely, input gate and output gate together with a memory cell and a forget gate to help retain important information throughout the model training phase.

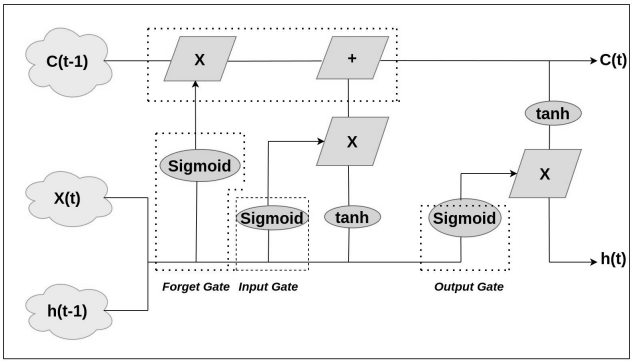

Fig. 4. LSTM Neural Network architecture.

TABLE II
LSTM MODEL ARCHITECTURE

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 304, 256)	264,192
lstm_1 (LSTM)	(None, 128)	197,120
flatten (Flatten)	(None, 128)	0
dropout (Dropout)	(None, 128)	0
dense (Dense)	(None, 32)	4,128
dense_1 (Dense)	(None, 1)	33

J. Hyperparameter Tuning

Hyperparameter refers to parameters passed to a machine learning model during the training phase of the model. Hyperparameter tuning is the process of using mathematical calculations to find the most effective set of values for the parameters that are passed to the models during the training phase. The values of these parameter are recursively tuned to generalise the understanding of the model on the training data. The hyperparameter selection for machine learning was done using Grid Search using the scikit-learn library in python. This method allows a single machine learning model to train on a different number of combinations of hyperparameters and returns the best possible combination having the highest accuracy. In LSTM model, the hyperparameter selection was done using random search and techniques early stopping and model checkpoints along with dropout layers were used to prevent overfitting during model training.

V. RESULTS

This study aims to introduce a comparative analysis of multiple models that are trained to predict the probability of a given consumer to repay the loan disbursed to them given their historical and current financial credit data. This study aims to enhance the use of such models in real-world scenario and reduce the manual effort it takes to go through the process of loan disbursing and streamline the process. The models discussed in this paper are trained on two categories of datasets: standard cleaned dataset having 303 features and a smaller dataset having top 100 features from the original dataset after applying Principal Component Analysis technique.

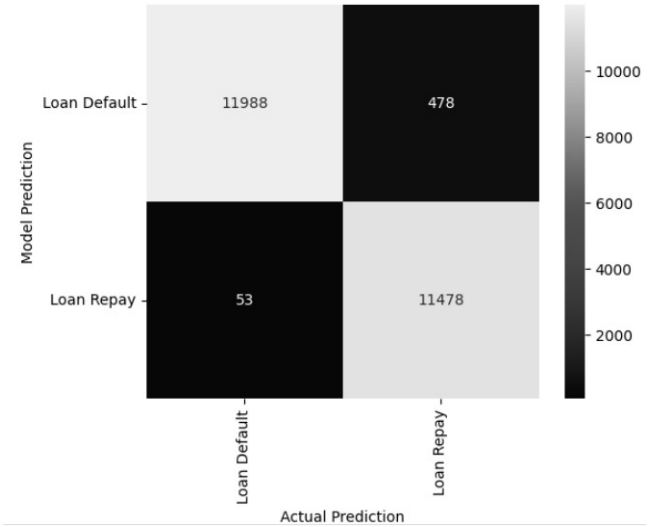

Fig. 5. Confusion Matrix for XGBoost model

The methodology used for evaluating the models is accuracy score. It is a simple model evaluation metric that calculates percentage of correctly classified values from the input and sum of all the inputs classified by the model.

$$\text{Accuracy} = \frac{(\text{T}P + \text{T}N)}{(\text{T}P + \text{F}P + \text{T}N + \text{F}N)} \tag{7}$$

Out of the models experimented, gradient boosting models performed the best by achieving the highest accuracy and fastest training times possible, closely followed by random forest and decision tree. K-Nearest Neighbours and probabilistic algorithms performed similarly on model evaluation. LSTM model performed well as it understood the current and historical records of the clients by retaining the information but as all deep learning models, it took the longest time to train on 10 epochs.

VI. CONCLUSION AND FUTURE WORK

The banking sector has supported entire economies through the process of disbursing loans to the people. Granting loans is one the major sources of income for a bank, hence, accurately anticipating if a given customer will repay or default the loan

TABLE III
PERFORMANCE COMPARISON BETWEEN BEFORE PCA AND AFTER PCA

Models	Before PCA		After PCA	
	Accuracy	Time (s)	Accuracy	Time (s)
XGBoost	98.52%	4.74	97.70%	0.56
AdaBoost	97.99%	36.75	98.10%	4.56
LightGBM	97.30%	2.95	97.86%	1.50
Random Forest	96.62%	7.27	96.18%	18.54
Decision Tree	95.12%	6.68	96.20%	0.70
Gaussian Naive Bayes	93.89%	0.26	98.09%	2.84
K-Nearest neighbors	93.89%	0.30	98.09%	3.84
LSTM	95.49%	671	96.00%	622

becomes a vital task. By making use of Machine Learning algorithms complex problems like the afore mentioned can be automated. Machine learning algorithms when fed with correct and enough amount of data can make up to 100% correct predictions.

In our experimental study, the performance of 8 different ML algorithms is compared using accuracy as metric for evaluation. Given the nature of dataset used, gradient boosting algorithms performed the best. XGBoost and AdaBoost models performed the best (reaching almost 99%) before and after applying PCA respectively. This study showcases that machine learning models can be applied on real-world banking data to automate the process of predicting loan defaults and solve the complex problem revolutionising the banking industry forever. Future research on this topic has the potential to scale these models to be practically applied in a real-world banking institution. Using different validation techniques and diversifying model choices, the accuracy can be improved further. In accordance with the results of this experimental study, it can be validated that machine learning and deep learning models hold significant credibility and potential in the financial sector.

REFERENCES

[1] RB, A., KR, S. K. (2021). Credit card fraud detection using artificial neural network. In Global Transitions Proceedings (Vol. 2, Issue 1, pp. 35–41). Elsevier BV. https://doi.org/10.1016/j.gltp.2021.01.006

[2] Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. In Procedia Computer Science (Vol. 162, pp. 503–513). Elsevier BV. https://doi.org/10.1016/j.procs.2019.12.017

[3] Awoyemi, J. O., Adetunmbi, A. O., Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI). 2017 International Conference on Computing Networking and Informatics (ICCNI). IEEE. https://doi.org/10.1109/iccni.2017.8123782

[4] Gao, J., Sun, W., Sui, X. (2021). Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. In A. Farouk (Ed.), Discrete Dynamics in Nature and Society (Vol. 2021, pp. 1–13). Hindawi Limited. https://doi.org/10.1155/2021/5080472

[5] Kandappan, V. A., Rekha, A. G. (2021). Machine Learning in Finance: Towards Online Prediction of Loan Defaults Using Sequential Data with LSTMs. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 2 (pp. 53-62). Singapore: Springer Singapore.

[6] Jain, A., Gupta, S., Narula, M. S. Loan Default Risk Assessment using Supervised Learning.

[7] Kibria, M. G., Sevkli, M. (2021). Application of deep learning for credit card approval: A comparison with two machine learning techniques. International Journal of Machine Learning and Computing, 11(4), 286-290.

[8] Zhang, H., Bi, Y., Jiang, W., Luo, C., Cao, S., Guo, P., Zhang, J. (2020). Application of random forest classifier in loan default forecast. In Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part III 6 (pp. 410-420). Springer Singapore.

[9] Patel, B., Patil, H., Hembram, J., Jaswal, S. (2020). Loan Default Forecasting using Data Mining. In 2020 International Conference for Emerging Technology (INCET). 2020 International Conference for Emerging Technology (INCET). IEEE. https://doi.org/10.1109/incet49848.2020.9154100

[10] Mohmad, Y. A. (2022). Credit Card Fraud Detection Using LSTM Algorithm. In Wasit Journal of Computer and Mathematics Science (Vol. 1, Issue 3, pp. 26–35). Wasit University. https://doi.org/10.31185/wjcm.60

[11] Shivanna, A., Agrawal, D. P. (2020). Prediction of Defaulters using Machine Learning on Azure ML. In 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE. https://doi.org/10.1109/iemcon51383.2020.9284884

[12] Yu, Y. (2020). The Application of Machine Learning Algorithms in Credit Card Default Prediction. In 2020 International Conference on Computing and Data Science (CDS). 2020 International Conference on Computing and Data Science (CDS). IEEE. https://doi.org/10.1109/cds49703.2020.00050

[13] Arora, S., Bindra, S., Singh, S., Kumar Nassa, V. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. In Materials Today: Proceedings (Vol. 51, pp. 110–117). Elsevier BV. https://doi.org/10.1016/j.matpr.2021.04.588

[14] Ndayisenga, T. (2021). Bank loan approval prediction using machine learning techniques (Doctoral dissertation).

[15] Madaan, M., Kumar, A., Keshri, C., Jain, R., Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, Issue 1, p. 012042). IOP Publishing. https://doi.org/10.1088/1757-899x/1022/1/012042

[16] Hassan, M. M., Mirza, T. (2020). Credit card default prediction using artificial neural networks. GIS Science Journal, 7, 383-390.

[17] Bacova, A., Babic, F. (2021). Predictive Analytics for Default of Credit Card Clients. In 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE. https://doi.org/10.1109/sami50585.2021.9378671

[18] Lai, L. (2020). Loan Default Prediction with Machine Learning Techniques. In 2020 International Conference on Computer Communication and Network Security (CCNS). 2020 International Conference on Computer Communication and Network Security (CCNS). IEEE. https://doi.org/10.1109/ccns50731.2020.00009

[19] Barbaglia, L., Manzan, S., Tosetti, E. (2021). Forecasting Loan Default in Europe with Machine Learning. In Journal of Financial Econometrics (Vol. 21, Issue 2, pp. 569–596). Oxford University Press (OUP). https://doi.org/10.1093/jjfinec/nbab010

[20] Kim, H., Cho, H., Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review. In Sustainability (Vol. 12, Issue 16, p. 6325). MDPI AG. https://doi.org/10.3390/su12166325

[21] Agarwal, A., Rana, A., Gupta, K., Verma, N. (2020). A Comparative Study and enhancement of classification techniques using Principal Component Analysis for credit card dataset. In 2020 International Conference on Intelligent Engineering and Management (ICIEM). 2020 International Conference on Intelligent Engineering and Management (ICIEM). IEEE.

https://doi.org/10.1109/iciem48762.2020.9160230
[22] Sheikh, M. A., Goel, A. K., Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE. https://doi.org/10.1109/icesc48915.2020.9155614
[23] Sheikh, M. A., Goel, A. K., Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE. https://doi.org/10.1109/icesc48915.2020.9155614
[24] Hasan, Basna Mohammed Salih, and Adnan Mohsin Abdulazeez. "A review of principal component analysis algorithm for dimensionality reduction." Journal of Soft Computing and Data Mining 2.1 (2021): 20-30.
[25] Anowar, F., Sadaoui, S., Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). In Computer Science Review (Vol. 40, p. 100378). Elsevier BV. https://doi.org/10.1016/j.cosrev.2021.100378
[26] Reris, R., Brooks, J. P. (2015). Principal Component Analysis and Optimization: A Tutorial. In Operations Research and Computing: Algorithms and Software for Analytics (pp. 212–225). INFORMS. https://doi.org/10.1287/ics.2015.0016
[27] Bhadula*, S., Sharma, S., Juyal, P., Kulshrestha, C. (2019). Machine Learning Algorithms based Skin Disease Detection. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 2, pp. 4044–4049). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijitee.b7686.129219
[28] Kamal, M., Pratap, A. R., Naved, M., Zamani, A. S., Nancy, P., Ritonga, M., Shukla, S. K., Sammy, F. (2022). Machine Learning and Image Processing Enabled Evolutionary Framework for Brain MRI Analysis for Alzheimer's Disease Detection. In D. Koundal (Ed.), Computational Intelligence and Neuroscience (Vol. 2022, pp. 1–8). Hindawi Limited. https://doi.org/10.1155/2022/5261942
[29] Bordoloi, D., Singh, V., Sanober, S., Buhari, S. M., Ujjan, J. A., Boddu, R. (2022). Deep Learning in Healthcare System for Quality of Service. In B. Nagaraj (Ed.), Journal of Healthcare Engineering (Vol. 2022, pp. 1–11). Hindawi Limited. https://doi.org/10.1155/2022/8169203