Home Credit Loan Default Prediction using Machine learning and Deep learning techniques

1st Rameshwari Kapoor
B.Tech (CSE)
Graphic Era Hill University
Dehradun, Uttarakhand, India
rameshwari.kapoor09@gmail.com

2nd Nilesh Bhanot
B.Tech (CSE)
Graphic Era Hill University
Dehradun, Uttarakhand, India
nileshbhanot18@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 adequate data and resources machine learning models can be trained to perform the same task reducing time and effort 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. Ensuring that the borrower will be able to pay back the loan on the proposed terms is the biggest challenge faced by banks in today's ever-changing world. This paper proposes to leverage the use of Machine Learning (ML) algorithms to predict which clients are more likely to default on their loans using historical financial data like credit and debit card data, bank account 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; deep learning;**

I. Introduction

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 repaid according to the accepted 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 (ML) and Deep Learning (DL) algorithms to predict the likelihood of loan default. These algorithms include Random Forest, Decision Tree, K- Nearest Neighbours, Gaussian Naïve Bayes, XGBoost, Light Gradient Boosting Machine (LGBM) , Ada Boost and Long Short Term Memory(LSTM). Use of machine learning methods 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 cleaned dataset to be fed to a machine learning model. We compare the performance of 8 ML and DL algorithms, out of which gradient boosting machines achieve nearly 99% accuracy, thus showcasing the promise of scaling these models to a real-world use case in the banking industry.

Identify applicable funding agency here. If none, delete this.

II. Literature Review

Asha RB et al. in [11] compared three algorithms named Support Vector Machines (SVM), K-Nearest Neighbour (KNN) and Artificial Neural Network(ANN). The conclusion of [11] reveals that ANN performed the best with an accuracy of 99.92% followed by KNN and SVM. Lin Zhu et al. In [14] compared Random Forest, Decision Tree, Support Vector Machine (SVM) and Logistic Regression. The experiment shows that Random Forest outperformed other algorithms with an accuracy of 98% followed by Decision Tree (95%) and SVM (75%). John O. Awoyemi et al. in [10] used hybrid sampling to handle imbalanced data. In [10] three algorithms named Naïve Bayes, KNN and Logistic Regression have been compared. The conclusion of [10] reveals that KNN performed

the best with an accuracy of 97. 9% followed by Naïve Bayes (97.6%) and Logistic Regression (54%). Anushi Jain et al. in [13] compared five algorithms named Logistic Regression, Support Vector Machine (SVM), Random Forest, XG Boost and Artificial Neural Network (ANN). The conclusion of [13] reveals that Logistic Regression is the best model to predict Loan default with an accuracy of 88.89% followed by Random Forest (88.85%) and XG Boost (88.57%). Huannan Zhang et al. In [3] compared Random Forest, Decision Tree and Logistic Regression algorithms to showcase the application of Random Forest Classifier in Loan default forecast. The conclusion of [3] is that the Random Forest Algorithm (86%) exceeds the decision tree (80%) and logistic regression classification (80%). Bhoomi Patel et al. in [1] compared four algorithms named Logistic Regression, Gradient Boosting, CatBoost Classifier and Random Forest to predict loan default. In [1] CatBoost Classifier outperformed other algorithms. It has an accuracy of 84.045%, whereas the other algorithms were 14.963%, 84.035%, and 83.514% respectively. Yue Yu in [9] compared four algorithms named Logistic Regression, Random Forest, Decision Trees and AdaBoost. The result of [1] shows that Random Forest gave the best accuracy of 82.12%. Saurabh Arora et al. in [12] compared six algorithms named K-Nearest Neighbour (KNN), Decision Tree, Random Forest, Logistic Regression, Support Vector Machine (SVM) and Naïve Bayes. The conclusion of [12] reveals that SVM is the best model to predict credit card default with an accuracy of 82% followed by Logistic Regression (81%) and Random Forest (80%). Theoneste Ndayisenga in his [8] has mentioned the use of Logistic Regression, Decision Tree, Support Vector Machines, Random Forest, KNN, Gausian Naive Bayes, Gradient Boosting and XGBoost. The result of the analysis of these algorithms shows that Gradient Boosting (81%) is the best model to predict bank default followed by XGBoost (80%).Mehul Madaan et al. In [2], compared Random Forest and Decision Tree algorithms to predict loan default. The conclusion of [2] is that Random Forest with an accuracy of 80% outperformed Decision Tree algorithm that gave an accuracy of 73%. The dataset that they used had biased data. Alžbeta Bačová and František Babič in their [5] have used Random Forest, AdaBoost and XGBoost for predictive analysis for credit card default. The results of [5] showed that the performance of these algorithms was very similar. Lili Lai in [7] has compared AdaBoost, XGBoost, Random Forest, KNN and Multi-Layer Perceptron algorithms to predict loan default. The conclusion of [7] is that AdaBoost outperformed all the other algorithms followed by XGBoost. Abhishek Agarwal et al. in [4] have mentioned about Logistic Regression, Random Forest, Decision Trees, Naïve Bayes and KNN in [4]. The main motive of [4] is to compare measures between the original dataset before and after applying the Principal Component Analysis. The conclusion of [4] is that the accuracy of Logistic Regression was best in both the cases and Decision Tress was not affected much. Mohammad Ahmad Sheikh et al. In [6] have used Principal Component Analysis (PCA) to analyse its importance. The conclusion of [6] Is that the model is marginally better after applying PCA.

III. Methodology

In the recent 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 proposed methodology is explained using the flowchart below.

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 to train and test the model. 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 which converts string data into numerical data by assigning 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 numerical features of the dataset are scaled to have a mean of 0 and standard deviation of 1. The formula for standard scaling is:

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

where, x is the original value of the feature,
μ is the mean of the feature values, and
σ is the 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 an algorithm called Principal Component Analysis (PCA). PCA is a dimensionality reduction technique that is used to reduce the number of features in a dataset while preserving all or most of the essential information. 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.

Feature Name	Feature Description
credamount$_7 70A$	Loan amount or credit card limit
cntpmts24$_3 658933L$	Number of months with any incoming payment in last 24 months
bankruptcy$_h istory$	Bankruptcy history of the client
age	Age of the client
education$_l evel$	Education level of the client
credit$_s core$	Credit score of the client
employment$_h istory$	Employment history of the client
debt$_t o_i ncome_r atio$	Debt-to-income ratio of the client
income	Income of the client
payment$_h istory$	Payment history of the client
late$_p ayment_h istory$	Late payment history of the client

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when

Fig. 1. Example of a figure caption.

writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] 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

[2] 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

[3] Huannan Zhang, Yilin Bi, Wangdong Jiang1, Chuntian Luo, Shengjia Cao, Peng Guo and Jianjun Zhang (year) Application of Random Forest Classifier in Loan Default Forecast

[4] 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

[5] 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

[6] 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

[7] 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

[8] Theoneste Ndayisenga (year) Bank Loan Approval Prediction Using Machine Learning Techniques

[9] 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

[10] Bagga, S., Goyal, A., Gupta, N., Goyal, A. (2020). Credit Card Fraud Detection using Pipeling and Ensemble Learning. In Procedia Computer Science (Vol. 173, pp. 104–112). Elsevier BV. https://doi.org/10.1016/j.procs.2020.06.014

[11] 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

[12] 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